AEM Accepted Manuscript Posted Online 9 March 2018
Appl. Environ. Microbiol. doi:10.1128/AEM.02855-17

A decade of *Streptococcus thermophilus* phage evolution in an Irish dairy plant

Running title: *Streptococcus thermophilus* phage evolution

Katherine Lavelle[1,2], James Murphy[1], Brian Fitzgerald[1,2], Gabriele A. Lugli[3], Aldert Zomer[2], Horst Neve[4], Marco Ventura[3], Charles M. Franz[4], Christian Cambillau[1], Douwe van Sinderen[1,2*] and Jennifer Mahony[1,2*]

[1] School of Microbiology, University College Cork, T12 YT20 Cork, Ireland; [2] APC Microbiome Institute, University College Cork; [3] Laboratory of Probiogenomics, Department of Chemistry, Life Sciences and Environmental Sustainability, University of Parma, 43124 Parma, Italy; [4]Department of Microbiology and Biotechnology, Max Rubner-Institut, 24103 Kiel, Germany.

*Corresponding authors. Mailing address: School of Microbiology, University College Cork, Cork T12 YT20, Ireland.

Telephone: +353 21 490 2443/1365

Fax: +353 21 4903101

E-mail: j.mahony@ucc.ie; d.vansinderen@ucc.ie

Keywords: Bacteriophage, *Streptococcus*, dairy industry, receptor binding protein, genomics

Abstract

Phages of *Streptococcus thermophilus* present a major threat to the production of many fermented dairy products. To date, only a handful of studies have assessed the biodiversity of *S. thermophilus* phages in dairy fermentations. In order to develop strategies to limit phage predation in this important industrial environment, it is imperative that such studies are undertaken and that phage-host interactions of this species are better defined. The present study investigated the biodiversity and evolution of phages within an Irish dairy fermentation facility over an eleven year period. This resulted in the isolation of 17 genetically distinct phages, all of which belong to the so-called *cos* group. Evolution of phages within the factory appears to be influenced by phages from other dairy plants introduced into the factory for whey protein powder production. Modular exchange, primarily within the regions encoding lysogeny and replication functions, was the major observation among the phages isolated between 2006 and 2016. Furthermore, the genotype of the first isolate in 2006 was observed continuously across the following decade highlighting the ability of these phages to prevail in the factory setting for extended periods of time. The proteins responsible for host recognition were analysed and carbohydrate binding domains (CBDs) were identified in the distal tail (Dit), the baseplate proteins and the Tail-associated lysin (Tal) variable regions (VR1 and VR2) of many isolates. This consolidates the notion that *S. thermophilus* phages recognise a carbohydrate receptor on the cell surface of their host.

Importance

Dairy fermentations are consistently threatened by the presence of bacterial viruses (bacteriophages or phages), which may lead to a reduction in acidification rates or

even complete loss of the fermentate. These phages may persist in factories for long periods of time. The objective of the current study was to monitor the progression of phages infecting the dairy bacterium *Streptococcus thermophilus* over a period of eleven years in an Irish dairy plant so as to understand how these phages evolve. A focused analysis of the genomic region that encodes host recognition functions highlighted that these proteins harbour a variety of carbohydrate binding domains, which corroborates the notion that phages of *S. thermophilus* recognise carbohydrate receptors at the initial stages of the phage cycle.

77 **Introduction**

78 *Streptococcus thermophilus* is one of the most extensively employed commercial
79 starter cultures, being widely used in the manufacture of fermented milk products
80 such as yoghurt and various cheeses (1, 2). Phage infection of *S. thermophilus* starter
81 strains may result in incomplete or failed fermentations with considerable economic
82 consequences to the dairy industry. Analysis of phage-host interactions is essential in
83 order to derive a detailed understanding of how these problematic phages recognise
84 and infect their host bacteria as a means to prevent or limit phage-mediated problems
85 in the dairy fermentation setting. The initial interaction between *S. thermophilus*
86 bacteriophages DT1 and MD2, and their hosts has been reported to involve three
87 phage proteins including the tail tape measure and the host specificity protein (3). The
88 receptor material for these phages is presumed to be a carbohydrate component of the
89 cell wall based on adsorption assays of *S. thermophilus* phages to differently treated
90 cell wall extracts of their hosts (4).
91 Several world-wide phage isolation studies have shown that two prevalent groups of
92 *S. thermophilus* phages exist. These groups are distinguished based on structural
93 protein content and their mode of packaging, which is determined by the recognition
94 of specific sequences, namely the *cos* and *pac* sites followed by specific or non-
95 specific cleavage of the DNA (5). The *cos* phages incorporate cohesive "sticky" ends
96 in their genomes, while *pac* phages employ a so-called headful DNA packaging
97 system and, therefore, may incorporate additional redundant DNA into their genomes.
98 PCR-based methods targeting the anti-receptor gene and the gene encoding the major
99 capsid protein, sometimes in combination with analysis of structural protein content
100 are two of the primary approaches currently used to identify to which of the sub-
101 groups new phage isolates belong (6-8). To date, 64 *S. thermophilus* phage genomes

have been sequenced, approximately half of which utilise the *cos* packaging mode (i.e. 34 of 64 sequenced phage genomes). Of the remaining 30 *S. thermophilus* phage genomes, 18 employ the headful packaging or *pac* method. In addition to the dominantly isolated *cos* and *pac* phages, two genetically distinct groups of *S. thermophilus* phages have recently been described. These include the 5093 group whose genomes bear greater similarity to prophages of non-dairy streptococci than those of dairy streptococcal phages (9), and the 987 group, whose genomes bear similarity to those of lactococcal P335 phages. The mode of DNA packaging of both of these newly described phage groups is not yet described (10). The identification of such novel groups of *S. thermophilus* phages highlights the importance of continued evaluation of phage biodiversity in dairy fermentation environments to identify resident populations and to develop robust starter strains and strain rotations.

Phage biodiversity surveys in dairy fermentation facilities are widely reported for lactococcal phages (11-18), while studies on *S. thermophilus* phage biodiversity are comparatively limited (9, 10, 19-21). While these studies have considerable merit in identifying the biodiversity of phages at a given time point, they do not provide temporal insights into the prevalence, maintenance, evolution and diversification of genetic lineages of phages within the industrial setting. In 2009, a longitudinal study of the evolution of lactococcal lytic 936 group phages in a Canadian cheese factory (22) highlighted that certain genetic lineages are able to survive in the plant for over a year and that genetic diversification was observed between the phages that were isolated over a 9 year period. To our knowledge, no such longitudinal studies of *S. thermophilus* phages have been published to date.

The success of phages in the dairy environment may be attributed to many factors including their ability to adapt to host defence mechanisms and their innate resistance

to chemical and thermal treatments applied in the dairy industry (23, 24). Significant effort has been invested towards understanding phages' adaptive responses to host-encoded phage-resistance systems in *S. thermophilus*, particularly with respect to clustered regularly interspaced palindromic repeat (CRISPR) immune systems (25-27). Studies of phage adaptive responses to thermal and chemical treatments have demonstrated the increasing insensitivity of phages of *Lactococcus lactis*, *Lactobacillus delbrueckii* and *S. thermophilus* to such interventions, thus establishing the requirement for industrial strategies to overcome this issue (24, 28, 29).

In the current study, 17 genetically distinct phages were isolated from an Irish dairy fermentation facility (factory A) over a period of more than a decade (2006-2016). The phages were isolated from cheese whey samples produced within the factory. Furthermore, cheese whey (as potential reservoirs of novel phage lineages) that had been introduced into a remote location on the factory site from other producers were also assessed for the presence of phages. The externally-derived whey is used in the production of whey protein powders and such whey protein powders have been demonstrated to be a rich source of dairy phages (30). Furthermore, phages display significant stability in this format providing an element of risk if whey protein powder is produced at the same site as the primary fermentation. The genomes of these phage isolates were sequenced and comparative genome analysis revealed two genetic lineages of *cos* type phages within the cheese production samples. Additionally, a further two genetic lineages exist among phage isolates associated with cheese whey acquired from other cheese producing facilities that are introduced into the plant for whey protein powder production. Focused analysis of the region encoding the predicted host interacting functions such as the distal tail (Dit), baseplate (BPP) and the host specificity protein/tail-associated lysin (Tal), highlighted the diversification

of these components through the acquisition of carbohydrate binding domains, which corroborates current thinking that phages of *S. thermophilus* recognise a carbohydrate receptor on their host cell surface.

Results

Phage isolations

Between 2006 and 2016, more than one thousand cheese whey samples from factory A were tested against *S. thermophilus* P1. This strain is the primary *S. thermophilus* production strain used to produce a particular Irish hard cheese for a period of approximately three to four months of the year. From these samples, a single phage type (named STP1) was isolated in 2006. This was considered the starting point of the evolutionary mapping of phages of the production strain P1 in this study. In subsequent years, samples were analysed for phages and restriction profile analysis identified the continued presence of the STP1 type phage in the production samples (Table 1) throughout the testing period of a decade despite the fact that the strain is only in use for at most one third of the year. Phage isolates with (minor) modifications in their restriction profiles were considered for further testing by host range analysis, multiplex PCR-based typing and genome sequencing. In addition to studying the phages from the cheese production facility of factory A, this study was aimed at identifying possible sources of different *S. thermophilus* phage genetic lineages that contribute to the development and evolution of phages in this particular cheese factory. Therefore, cheese whey samples (2,043 samples) acquired from other factories for whey protein powder production were also tested for the presence of phages of *S. thermophilus* and their relatedness to those identified in the production plant was assessed. These samples were tested against a panel of 52 *S. thermophilus* dairy strains that were available within our collection to obtain maximum phage diversity.

In this study, 17 genetically distinct phages were isolated, eight of which were isolated from factory A-derived cheese whey samples in 2006, 2008 and 2015 (STP1 in 2006, STP2 in 2008, and A0, B0, C0, 9B4, 16B8 and 31B4 in 2015). Between 2008 and 2015, over one hundred phage isolates with near identical restriction fragment length polymorphism profiles to those of STP1 (and on occasion STP2) were identified only within the cheese factory itself highlighting the dominant application of particular starter cultures and the enduring nature of the STP1/STP2 phages in the plant (Table 1). While no novel isolates were observed in the factory during this seven-year period (2008-2015), a novel phage genotype (B5) was isolated in the external whey samples in 2012 and additional novel genotypes were identified in the external whey samples in 2014 (MM25 and M19), 2015 (9A, L5A1, 7A5, 7T) and 2016 (V2 and R1) (See phage list in Table 2). A noteworthy point is the spread of samples across the eleven year period. While an approximately equal number of samples from the factory were tested each year, the number of externally derived samples was increased considerably from 2012 onwards to expand the potential for isolation of novel genotypes. Thus from 2006 to 2011, approximately 100 externally derived samples per year were tested, while from 2012 to 2016 approximately 300 samples were assessed. This resulted in an increase in the number of genotypes identified highlighting the benefit of extensive screening. 63 % of samples from factory A were positive for the presence of phages targeting strain P1 with titres ranging from 10^2-10^5 pfu.ml^{-1} in the original whey samples. 68 % of the externally derived samples were phage positive against at least one of the 52 strains included in the testing panel with phage titres ranging from 10^2-10^8 pfu.ml^{-1}. All phage isolates were identified as *cos*-type phages by multiplex PCR with the corresponding PCR product of 170 bp visualised on a 1 % agarose gel (data not shown).

Host range analysis

All samples derived from the factory were initially tested against *S. thermophilus* P1 (the strain used in production), and phage isolates were propagated on this strain and subsequently used in a challenge against a collection of 51 additional *S. thermophilus* strains. The majority of phage isolates from the factory were identified to have a common secondary host (P2), while phage-specific infection profiles were also observed (Table 2). In order to assess the extent of phage biodiversity, whey samples acquired from other factories were screened against the panel of 52 *S. thermophilus* strains without a pre-screening on P1 alone (See methods). Many of these samples also contained phages capable of infecting *S. thermophilus* P1 with six of the nine phages isolated from externally derived whey identified originally on this strain only and upon production of high titre lysates ($>10^7$ pfu.ml^{-1}), additional hosts were identified (Table 2). Interestingly, phages MM25, M19 and R1, which were identified on primary hosts other than P1 in the phage screen (Table 1), have unique and narrow host range profiles. While these phages were not isolated on strain P1, when a high titre lysate of MM25 was produced on its primary host strain, a very small subpopulation was capable of infecting P1 (at an efficiency of plaquing of 10^{-6}) highlighting their ability to readily adapt to this strain (and others).

Phage lineages

During the eleven year period between 2006 and 2016, more than 300 individual phage isolates were compared (by restriction profiling, data not shown), of which the

vast majority (96 %) was identified as exhibiting STP1-like profiles (lineage 1) in both the factory and externally derived whey (Tables 1 & 2). However, a small number of isolates were identified for which clearly distinct restriction profiles and/or host ranges were observed. These were considered for further analysis resulting in the identification of 17 distinct *S. thermophilus* phages, which were then subjected to whole genome analysis (Supplementary Table S1). Based on phylogenetic analysis of the overall nucleotide sequences, there appears to be four genetic lineages of phages, including two derived from phages within the cheese production facility of factory A (lineages 1 & 2) and two from externally derived whey samples only (lineages 3 & 4) (Fig. 1). The first lineage is that of STP1, the first isolate in 2006, further including phage isolates STP2, A0, B0, C0, L5A1, B5 and 7A5. Phages of this lineage appear to share a similar host range profile and are characterised by their ability to primarily infect strains P1 and P2 (Table 2). Five of these phages originate from the cheese whey from factory A, while three were isolated from cheese whey from external sources (B5 in 2012, and L5A1 and 7A5 in 2015) (Fig. 1 & Table 1). In order to perform a more focused analysis of the genetic content of the 17 phage isolates, the most closely related isolates were identified by alignment of their nucleotide sequences (Fig. 1). This formed the basis of a neighbour mapping of phage genomes following the order identified in the phylogenetic analysis as displayed in Figures 2 (lineage 1 and 2 phage isolates) and 3 (lineage 3 and 4 phages that have not yet been observed in factory A). In this analysis, STP1 was the base comparator as it was the first isolate and it served to highlight the major regions of divergence between phages of different lineages and within lineage 1 phages.

While lineage 1 phages appear highly related, some insertions/deletions and minor rearrangements within the lysogeny and replication modules are observed in

comparison to the first isolate, STP1 (Fig. 2). The genome of STP2, isolated in 2008 was essentially identical to that of STP1 with some point mutations throughout the genome and minor sequence variations at the genomic termini. The remaining lineage 1 phages display a high degree of similarity (Fig. 2). Therefore, minor deletions and genetic rearrangements appear to be the major feature of the evolution of lineage 1 phages, while localised genetic acquisitions were also observed in L5A1, 7A5 and B5 in the lysogeny- and replication related modules (Fig. 2). It is possible that other companies also use *S. thermophilus* P1 or closely related strains thereby explaining the prevalence of this phage lineage in samples derived from external sources (Table 1).

Lineage 2 is represented by phages related to 31B4 and further represented by phage isolates 7T, 9B4, 16B8, V2 and R1 (Fig. 1). These phages appear to be highly similar to the STP1 lineage (lineage 1), most notably in the genomic region encoding the structural components, while having acquired a distinct genomic region within the predicted lysogeny and replication gene modules with only the very rightward end of the genome displaying similarity between B0 (lineage 1) and 31B4 (lineage 2) (Fig. 2). It seems highly likely that the phages isolated in the factory (31B4, 9B4 and 16B8) have recombined with (one of the) externally derived phages such as 7T, as the latter phage was isolated in the externally-derived whey prior to the identification of similar restriction profiles and genetically similar isolates in the factory later the same year, possessing an almost identical replication module (Fig. 2). Phage 7T appears to be the ancestor (or its closest relative) of this lineage since it was the first of this type to be isolated in early 2015 (Table 3) with additional members of this lineage isolated both in the factory- and externally derived whey samples appearing later in 2015 and 2016, respectively (Table 1).

Lineage 3 consists of MM25 and 9A, both of which were isolated from externally derived whey samples. Lineage 3 contains no members directly derived from factory A. Similarly, lineage 4, which has only one constituent member (M19), was isolated from an externally derived whey sample in 2014. The genetic region encoding the structural elements of phages of lineages 3 and 4 bears significant similarity to those of lineages 1 and 2 (Fig. 3 displays the comparison of lineage 3 and 4 phage genomes to each other and to that of the lineage 1 phage STP1); however, these phages are considerably divergent in their putative replication and lysogeny modules although there is moderate similarity between lineage 1 and 4 phage isolates in the replication region suggesting a possible shared ancestry (Fig. 3). Interestingly, the primary host of MM25 and M19 is not P1 indicating that these phages are derived from factories that likely do not employ this strain although they have demonstrated the ability to adapt to infect P1 (at a frequency of ~10^{-6}), among other strains. Conversely, 9A was initially isolated on strain P1 and exhibits a host range similar to those of a number of lineage 1 and 2 phages (Table 2). This may indicate that 9A was exposed to strain P1 earlier than MM25 thus providing the phage with the opportunity to adapt. Indeed, phage 9A may have appeared in the factory prior to its first detection in this study (in 2015). The abundance of 9A (and MM25 and M19) may have been below the detection threshold of this study thus precluding its' isolation in previous sampling years.. Given the distinct host ranges of these isolates, it is noteworthy that lineage 3 and 4 phages display reduced similarity in the genomic region encoding host recognition functions (Fig. 3). Furthermore, these three isolates (MM25, 9A and M19) were the only members of their kind identified in this study highlighting the low incidence of these phage genotypes.

Genome organisation

The genomes of phage isolates sequenced in this study are similarly organised with four identifiable functional modules based on BLASTP analysis, i.e. the structural/packaging, lysis, lysogenic and replication modules (Fig. 2 and 3). The genomes of the isolated phages are 34.0 - 36.8 kb in length and carry between 39 and 48 predicted ORFs (Table 2). BLASTN analysis of STP1 highlighted that this phage bears most significant similarity to the *cos*-type phages Abc2 and DT1 with almost identical sequences shared across approximately 70 % of their respective genomes. Three major functional modules are observed in the genomes of the isolated phages associated with morphogenesis, lysogeny and replication, and these are highlighted below in more detail. The major region of divergence is contained within the lysogeny and replication modules as exemplified in Figures 2 and 3.

The most leftward functional module on the phage genomes (as depicted in Figs. 2 & 3) encodes the predicted structural components and DNA packaging system. Within the structural module of the bioinformatically analysed phage genomes, it is possible to identify the small and large terminase-encoding genes (*terS* and *terL*), as well as genes specifying the portal, scaffolding, major capsid, head-tail joining, major tail, tail tape measure and baseplate proteins. Pfam searches with the tail tape measure protein (TMP) sequence revealed two domains at the carboxy-terminus, namely a Cysteine histidine-dependent amidohydrolase/peptidase (CHAP) and a Soluble Lytic Transglycosylase (SLT) domain that resemble lytic domains often associated with tail-associated lysins of phages (34, 35). The majority of the structural protein-coding regions are highly conserved between all isolates with some notable exceptions. Among these, the so-called host specificity proteins encoded by the phage isolates show significant divergence between phages belonging to lineages 1/2 and 3/4 (Figs.

2 - 4), which is likely a reflection of the distinct primary hosts on which these phages were isolated and the different sources of isolation of the phages.

The second functional module is represented by the lysis cassette, which typically encompasses holin and lysin-encoding genes required for progeny phage release at the end of the phage cycle. The vast majority of *S. thermophilus* phages reported to date encode two lysins and the phages isolated in this study are no exception to this. The first of the two lysins encoded by these phages possesses an amidase/peptidoglycan hydrolase domain and would be expected to encode a functional lysin. Additionally, the second (where identified) bears similarity to the lysin encoded by the *S. thermophilus* phages Abc2 and ALQ13.2 for which genome sequence data is available. The stacking of genes involved in lysis may imply that one of the lysins is non-functional through mutation or deletion events or additional genes encoding endopeptidases/lysins were acquired through homologous recombination with other streptococcal phages.

The third gene cassette relates to lysogeny functions with an identifiable repressor for the lytic and lysogenic cycles while a number of genes encoding proteins of unknown function were identified in several of the phage isolates as well (Fig. 2 & 3). The suggestion that this genomic region is a recombination hot-spot may explain the persistence of this genomic region in virulent phages (36). Indeed, in the present study, the lysogeny module is among the most divergent genomic regions within the analysed genomes. The genomic location of the lysogeny cassette is consistent with that of previously sequenced *S. thermophilus* phages, i.e. between the lysis and replication modules (37, 38).

The fourth and last module encodes replication-associated proteins, such as DnaC and a single-stranded DNA binding protein. This module is among the most divergent

regions of the genomes among the 17 phage isolates (Fig. 3). The observed diversity among the replication regions of these phages is perhaps the most distinctive feature of the four lineages of phages isolated in this study. Lineages 3 and 4 phages display completely distinct replication modules to each other (Fig. 3) while that of M19 (lineage 4) bears some observable relationship to lineage 1 phages such as STP1 (Fig. 3). Furthermore, the replication modules of lineage 1 and 2 phages appear to be distinctive with very limited similarity in this region as exemplified by B0 and 31B4 in Fig. 2. This highlights that among the *cos* phages isolated in this study, the most significant genomic source of diversity is within the replication module.

Tail tip functions harbour carbohydrate-binding domains

In *Siphoviridae*, the first gene following the TMP-encoding gene is the distal tail protein (Dit)-encoding gene (39), which in turn is typically followed by the gene specifying the tail-associated lysin or Tal protein. The Tal protein may encompass several functions, which may or may not include lytic activity, but in which the ~400 N-terminal residues have the same topology among many *Siphoviridae* phages (39). The TMP of the phages that are subject of the current study is, as mentioned above, predicted to contain lytic activity. The Tal protein encoded by the phages studied in the present study contains, in addition to the topologically conserved N-terminal portion, a previously described host-specificity region or receptor binding protein (RBP) domain (3, 40). We will therefore refer to this protein as Tal-RBP to represent the apparent dual function of this protein. Finally, the ORF downstream of the Tal-RBP-encoding gene, is coined here as *bpp*, based on the notion that it encodes a baseplate protein (BPP). Several ORFs in the DNA region encoding tail components have been found to contribute to sugar recognition in *Siphoviridae* phages (41-46),

and thus the sequences of Dit, Tal-RBP and BPP were analysed for potential carbohydrate (or other) binding domains.

Dit proteins have been designated to be either “classical”, i.e. with a short sequence, or “evolved”, i.e. bearing carbohydrate binding domains (CBDs) (11, 43). The phages in this study all encode evolved Dit proteins since they harbour a CBD that shares structural similarity (HHpred score 99.9 %) with the CBD of the evolved Dit from *Lactobacillus casei* phage J-1 (43).

The Tal-RBPs encoded by the phages isolated in this study range in size from 906 aa (MM25) to 1114 aa (9A) (Fig. 4). This size range is consistent with previously studied *S. thermophilus* phage Tal proteins (coined host specificity determinants in (40)). The Tal-RBPs of *S. thermophilus* phages are characterised by the presence of a conserved N-terminal region (Tal domain; 1 to ~400) followed by one or two variable regions, named VR1 and VR2, flanked by multiple collagen repeats (Fig. 4) (40). Sequence analysis of Tal-RBPs of the phages isolated in this study revealed that each previously termed variable region is in most cases a predicted CBD and the variation in size of the Tal-RBP proteins may be accounted for by the presence or absence of such CBDs.

MM25 and M19 encode the smallest Tal-RBPs of 906 aa and these proteins lack a VR1 region, while at aa positions 488-570 a partial 5E7T_B domain, which corresponds to the recently described CBD of a lactococcal phage accessory baseplate protein called BppA (44), is observed. This domain is followed by a CBM4 family Igu1_A domain at positions 570-715 based on HHPred analysis in MM25 (Fig. 4). Together, these domains constitute the VR2 domain. In addition, the Tal-RBPs of these phages did not harbour any obvious collagen repeat motifs (G-X-Y). The host-recognising VR2 regions of MM25 and M19 were distinct from one another, which is consistent with their unique host range profiles. Phages belonging to lineages 1 & 2

described above encode Tal-RBPs that harbour a 5E7T_B (BppA) domain with greater than 97 % probabilityThe presence of the 5E7T_B BppA-like CBD in the Tal-RBPs of lineage 1 & 2 phages (Fig. 4) in this study correlates with the position of the first variable region, VR1, described in phages DT1 and DT2 (40). Figure 4 highlights the phylogeny of the Tal-RBPs of the phages isolated in this study and further illustrates the presence of CBDs and the relative size of these proteins in representative lineage 1/2 (STP1), 3 (9A) and 4 (MM25) phage isolates. To further assess if VR1 is typically representative of the BppA CBD in other *S. thermophilus* phages, the Tal-RBPs of DT1 and DT2 were analysed using HHpred revealing the presence of a similar CBD in the same relative position in both Tal-RBPs. This indicates that VR1 is, in fact, a variably present, (apparently) non-essential, yet possibly accessory BppA-like CBD. In addition, the Tal-RBP of DT2 harbours a second CBD (now known to be a 5E7T_B BppA-like domain with 94 % probability) covering part of the position formerly identified as VR2. Similarly, the Tal-RBP of phage 9A harbours two adjacent BppA-like CBDs akin to DT2 and is the largest among the predicted Tal-RBPs encoded by phages in the present study at 1114 aa, in agreement with the acquisition of multiple CBDs (Fig. 4). The CBDs in the Tal-RBPs of the phages from the present study are also flanked by variable numbers of collagen repeat motifs (G-X-Y). Therefore, it is now clear that the variable regions are largely represented by CBDs and it is possible that the repeat motifs may aid the recombination and insertion of such domains. The VR2 domain includes the host recognition domain (as previously defined) and may additionally include a CBD in some cases giving rise to the size variation among these proteins (and perhaps the ability to bind to a range of carbohydrate motifs).

In addition to the classical Tal-RBPs (previously termed the host specificity protein) of *S. thermophilus* phages, their genomes also typically harbour a gene downstream of the Tal-RBP-encoding gene whose product is currently of unknown function although it is implicated in host interactions. HHPred analysis of this protein in this phage family (e.g. DT1 as a representative of the group of isolates) highlighted the presence of a C-terminal domain with predicted structural similarity to the phage RBP of TP901-1 (97 % probability), which is involved in *L. lactis* cell wall polysaccharide binding. This strongly supports the notion that this highly conserved protein is part of the tail tip and involved in host interactions.

Morphological analysis

All phage isolates were identified as *cos*-type phages related to the well described phage *cos*-type phage DT1, which exhibits a long non-contractile tail and an isometric head. To confirm that the phages isolated in this study conform to the expected morphological characteristics of these phages, STP1 and MM25 were (randomly) selected for characterisation by electron microscopy. STP1 and MM25 were found to possess tails of 253.9 ± 9.7 nm (n=19) and 247.6 ± 6.2 nm (n=24), respectively and heads with diameters of 56.4 ± 1.6 nm (n=20) and 55.7 ± 2.1 nm (n=24), respectively (Fig. 5). Given the significant sequence relatedness of the majority of isolates in this study, it is likely that all these phages are representative of the overall group of isolates. Interestingly, a tail-associated, “feather-like” appendage protruding from the tail tip was observed for both phages of 46.7 ± 2.3 nm (n=20) (STP1) and 45.9 ± 2.9 nm (n=23) (MM25) including a bridging fibre structure of approximately 12 nm (n=20).

475

476 **Discussion**

477 Limited studies pertaining to the diversity, persistence and evolution of *S.*
478 *thermophilus* phages in dairy fermentation plants have been performed (6, 7, 10, 20,
479 21, 32, 33, 47). Among those that have been performed, limitations such as
480 sequencing of very localised regions of the phage genomes such as the VR2 regions
481 (48) and sequencing of limited numbers of phage isolates (36) have caused knowledge
482 of the phage-host interactions in *S. thermophilus* to lag behind those of their
483 lactococcal counterparts. However, these studies have been very useful in
484 demonstrating that *cos* phages predominate in dairy samples where *S. thermophilus*
485 starter strains/cultures are employed. Furthermore, it has also been demonstrated that
486 phages are abundant in whey protein powder (30) and that, in particular, *S.*
487 *thermophilus* phages are synonymous with modular exchange in terms of their
488 evolutionary development pathways (10, 38, 49). The theory of modular exchange of
489 *S. thermophilus* phage genomes was clearly demonstrated between the classical *cos*
490 phages Sfi21 and 7201, among others (36). This highlights that while slow
491 evolutionary drift may occur in this species, homologous recombination of small sets
492 of genes or, indeed, entire modules such as the replication module in the case of the
493 above-mentioned phages (36) or the morphogenesis module in the case of the 987
494 phages (10) is a widely observed phenomenon. In the broader context, phage genome
495 evolution has been proposed to proceed via a high or low gene content flux depending
496 on the host species; the lytic or temperate nature of the phages and balance of these
497 two types of phages within a given species (50). In *S. thermophilus*, the incidence of
498 lysogeny is reportedly low (51) and therefore, it would perhaps be expected that these
499 phages would likely follow the low gene content flux as high gene content flux is

more typically observed among temperate phages. However, this does not appear to be the case for *S. thermophilus* phages where modular shuffling and recombination are observed among lytic phages.

In the present study, we report the genomic structure and suggest the evolutionary development of 17 novel *S. thermophilus cos* phages isolated from whey samples from a single Irish factory as influenced by whey that was imported into the factory site for processing into why protein powder. Given that whey protein powder is known to act as a rich reservoir for dairy phages, it is unsurprising perhaps that the phages from the externally-derived whey appear to have recombined with phages in the cheese factory itself (30). The identification of novel genotypes in the external wheys often coincided with the identification of similar genotypes in the factory indicating the likely entry of the phages from the imported whey into the cheese factory and/or recombination with the existent phage population. This is significant since there may be a potential for recombination with existing lytic phages in the factory or, to a lesser extent, integrated prophages of P1 (although it is currently not known if P1 harbours prophages in its genome) or other strains employed in the factory. While the incidence of lysogeny is not very high in *S. thermophilus*, modular exchange of lytic phages is a well-established phenomenon in phages of this species (10, 31-33). For example, phages 7A5 and L5A1 predate A0, B0 and C0 in the phage isolations during 2015 and there is considerable relatedness between these lineage 1 phages (Figs. 1 & 2). Furthermore, the lineage 2 phage 7T predates 9B4, 16B8 and 31B4 and the genomes of these phages were demonstrated to exhibit novel replication regions relative to lineage 1, 3 or 4 phages (Figs. 2 & 3). Interestingly, the identification of 7T in March 2015 in the externally-derived whey samples was

525 followed by the identification of phage isolates with similar DNA restriction profiles
526 in the factory (represented by 9B4, 16B8 and 31B4) later in 2015 (June). Thus, it is
527 likely that lineage 2 phages primarily derived from lineage 1 phages as their
528 packaging and morphogenesis modules are highly conserved and that homologous
529 recombination and modular exchange with 7T-like phages from the externally-derived
530 phages resulted in the presence of lineage 2 phages in the factory. Lineage 3 and 4
531 phages appear to exhibit a higher degree of genetic novelty and divergence and are,
532 therefore, considered distinct genetic lineages to those represented by lineages 1 & 2.
533 While lineage 3 and 4 isolates MM25 (2014), 9A (2015) and M19 (2014) were
534 characterised as harbouring significant diversity in their genetic content, there does
535 not appear to have been a coincidence of phages with similar genotypes in the factory
536 as yet. These phages were only isolated on two occasions (Table 1) and, therefore,
537 they are less abundant and prevalent in the processing site thereby limiting their
538 development and transfer to the cheese factory. However, their presence on the
539 factory site harks a warning to producers to monitor for the continued presence of
540 such phages that may become problematic if they become prevalent or more
541 abundant.
542 Longitudinal studies such as this one are important to understand the natural
543 evolutionary processes at play in the industrial context and reinforce the notion of
544 evolution of *S. thermophilus* phages by modular rearrangements and acquisitions
545 while minor evolutionary shifts are also observed (10, 31, 49). Furthermore, the
546 morphological analysis of phages STP1 and MM25 revealed the presence of a feather-
547 like tail appendage (Fig. 5). This feather-like appendage has been observed recently
548 for some *cos*-type phages (20, 33), although it is not widely described among this
549 group of phages. No obvious genetic element could be identified that is unique to

these phages compared to previously sequenced *S. thermophilus* phage genomes. However, early attempts at imaging CsCl-purified lysates of STP1 were largely unsuccessful due to instability of the phages and many separated tails and heads were observed, while in addition the feather-like appendage was not observed (data not shown). Therefore, it is possible that this appendage is a fragile structure that may be destroyed or removed by harsh treatments and/or ultracentrifugation. The preparation of fresh crude lysates that were subsequently diluted to remove media and contaminating background artefacts proved a more effective approach to the analysis of these phages that retain these delicate appendages. Knowledge of the presence of such structures that may play a role in host interactions is vital to developing a detailed understanding of the means by which phages of *S. thermophilus* recognise and attach to their bacterial hosts.

The Tal-RBP encoded by phages is the primary determinant of host recognition and attachment and in 2001, the modular arrangement of the host specificity proteins (Tal-RBPs) of seven *S. thermophilus* phages was analysed (40). In this study, the Tal-RBPs were determined as putative RBPs of the phages, and were characterised as presenting with up to three domains: (i) a conserved N-terminal region of 491 aa (domain 1); (ii) the VR1 region, which is present in some but not all Tal-RBPs (domain 2), and (iii) the VR2 which is involved in host recognition (domain 3). In the present study, detailed bioinformatic analyses revealed that the VR1 and VR2 regions are often represented by predicted carbohydrate-binding domains. This consolidates the notion that *S. thermophilus* phages recognise a carbohydrate surface receptor. Furthermore, the presence of a predicted cell wall polysaccharide or teichoic acid-interacting domain in the second baseplate protein (BPP) provides additional insights into the

complexity of the interactions of these phages. Structural bioinformatics is a very useful tool to shed light on viral "dark matter" and this is a noteworthy example of its application. Indeed, it may have implications for the interactions of phages of non-dairy streptococci that may employ similar receptor material and provides a route of investigation for these phages.

Lactococcal phage-host interactions have become a paradigm for Gram-positive bacteria and their infecting phages and this is warranted by the extensive application of lactococcal starter strains in the dairy industry and the persistence of their phages in the dairy fermentation setting (44, 52-54). Despite the industrial importance of *S. thermophilus* starter cultures in the dairy industry, the phage-host interactions of this species have not enjoyed as much attention as their lactococcal counterparts. Therefore, it is essential to generate data on the diversity and evolution of these phages, and to unravel the intricacies of their interactions with their respective host bacteria. The finding of presumed CBDs in the tail tip structural proteins of *S. thermophilus* phages provides clear direction for future studies relating to these phages and the means by which they recognise and attach to their cognate hosts to expand current knowledge on this industrially important subject. Future research will focus on the functional characterisation of individual proteins and protein complexes that harbour CBDs to define the role of the individual domains in generalised or specialised binding to the host. The identification of CBDs in phage structural proteins will provide a basis for the targeted analysis of different CBD types to assess the range of binding activities among such phages.

Materials & Methods

Bacterial strains, bacteriophages and media

Bacterial cultures were grown in M17 broth (Oxoid, Hampshire, U.K.) supplemented with 0.5 % lactose at 42 ˚C. Phages were isolated from dairy whey samples from an Irish cheese production facility (factory A), both those produced in-house as well as those obtained from cheese whey derived from other factories that were destined for whey protein powder production on the premises of factory A. Two phage screening approaches were employed in this study. In the first approach, all samples (>1000 samples) derived from the cheese factory (Factory A) were initially tested against *S. thermophilus* P1 using the double agar plaque assay method defined by Lillehaug (55) since this is the *S. thermophilus* strain that is primarily used in the cheese production process of factory A. In all cases, the same master stock of *S. thermophilus* P1 was used to generate the P1 culture for phage testing (and was sourced from the commercial starter supplier) to ensure that the culture had not changed across the testing period of 11 years. The second approach involved 2,043 whey samples derived from other factories (using unknown starter cultures), which were tested against a panel of 52 *S. thermophilus* strains obtained from the UCC strain collection (representing historical dairy isolates, Supplementary Table S2) to identify potential novel phage isolates that may be introduced into the plant. This second part of the study therefore reflects 106,236 (52 strains tested against 2,043 samples) sensitivity assays performed over the eleven year period. Phages were propagated on relevant hosts as indicated in Table 1 at 42 °C with the addition of 10 mM $CaCl_2$, filtered after lysis had occurred and stored at 4 °C until required.

Host range analysis

All bacteriophages were propagated to a titre of (at least) 10^7 pfu.ml^{-1} and their host ranges were subsequently assessed on 51 additional strains from various sources. Host range analysis was performed using the previously described spot test method (56). To verify the results of the spot assays, enumeration of the level of sensitivity of each strain was defined by plaque assays (55). All assays were performed in triplicate and the efficiency of plaquing (E.O.P.) was calculated and is reported in Table 2. The E.O.P. was defined as the ratio of the average titre of the phage on the test (secondary) host strain to the average titre of the primary propagating host strain.

Multiplex PCR

To define if isolated phages belong to the *cos*- or *pac*-type *S. thermophilus* phages, multiplex PCR was performed using phage DNA as template. The multiplex PCR was based on the method of Quiberoni and colleagues using the conserved sequences of the gene encoding the major capsid protein of seven *S. thermophilus* phages belonging to both the *cos* and *pac* subgroups (19). The primer sequences based on the *cos* phages were as follows: *cos*FOR (5'-ggttcacgtgtttatgaaaaatgg-3') and *cos*REV (5'-agcagaatcagcaagcaagctgtt-3') with an expected product size of 170 bp and those based on the *pac* phages were *pac*FOR (5'-gaagctatgcgtatgcaagt-3') and *pac*REV (5'-ttagggataagagtcaagtg-3') with an expected product size of 427 bp. The resulting amplicons were applied to a 1 % agarose gel and visualized by UV transillumination.

Phage DNA extraction

DNA for genome sequencing was extracted from 50 ml of fresh phage lysate (~10^8 pfu.ml^{-1}), which was first treated with 1 µg/ml DNAse and RNAse at 37 ˚C for 30 minutes. Following centrifugation at 13,200 x *g* for 15 min, the lysate was transferred

649 to a new tube, after which polyethylene glycol 8000 and NaCl were added to a final
650 concentration of 10 % and 0.5 M, respectively, and the resulting suspension was then
651 incubated at 4 ˚C overnight. Subsequently, the suspension was centrifuged at 17,700 x
652 *g* for 15 min and the supernatant removed. The PEG-precipitate was resuspended in 5
653 ml of TE buffer (pH 9.0) and treated with 120 µl of 20 mg/ml proteinase K for 20 min
654 at 56 ˚C. Potassium acetate was added to a final concentration of 1 M followed by
655 incubation on ice for 20 min before centrifugation at 13,200 x *g* for 10 min. The
656 supernatant was then phenol/chloroform-extracted (25:24:1
657 phenol:chloroform:isoamyl alcohol, Sigma Aldrich) (at least) twice and the aqueous
658 phase precipitated with 2.5 volumes of ice cold 96 % ethanol and 0.1 volume of 3M
659 sodium acetate (pH 4.8). Subsequent to centrifugation, the pellet was washed in 70 %
660 ethanol and resuspended in 100 µl of TE buffer (pH8.0).
661
662
663
664 **Genome sequencing, assembly & annotation**
665 Sequencing of the STP1 genome was conducted using a GS-FLX Titanium sequencer.
666 10 µg of DNA was extracted and verified by nanodrop quantification and
667 confirmatory PCR-based ID tests and restriction profile analysis were conducted on
668 the DNA extract prior to shipment to the contract sequencing facility (Agencourt
669 Bioscience, MA, USA). Chromosomal DNA was mechanically sheared via a
670 Hydroshear device (Genemachine, San Carlos, CA) and fragment size selected (3 kb)
671 on an agarose pulsed-field gel electrophoresis, excised and purified. A similar
672 approach was employed for the sequencing of STP2 but by a different service
673 provider (Macrogen, Seoul, Korea). The files generated by the 454 FLX instrument

were assembled with GSassembler (454 Lifesciences, Branford, CT, USA) to generate a consensus sequence. The assembly was entered into Staden (57) and additional sequencing walks were performed resulting in a single, gapless contig. The remainder of the genomes were sequenced using Illumina MiSeq technology (GenProbio, Parma, Italy). MIRA (mimicking intelligent read assembly) version 4.0.2 was used for *de novo* assembly of MiSeq-derived phage genome sequences to generate a consensus sequence. Open reading frames (ORFs) were predicted using a combination of Prodigal v2.6 and BLASTX (58, 59) followed by manual assessment, curation, and correction of the predicted ORFs. Functional annotations were generated using BLASTP (60) analysis against the non-redundant protein database (nr) provided by the National Centre for Biotechnology Information (located at: http://blast.ncbi.nlm.nih.gov/Blast.cgi) as well as using the MEGAnnotator pipeline (61). Proposed protein functions were validated by querying protein domain databases Pfam (62), the National Center for Biotechnology Information (NCBI) Conserved Domain Database (63), and by performing homology prediction searches using HHPred (64). The genomes were searched for the presence of potential transfer RNA (tRNA) genes using tRNAscan-SE (65). The genomic characteristics of the sequenced phage isolates are presented in Table 2. Quality improvement of the genome sequences involved sequencing of PCR products across the entire genome to ensure correct assembly, double stranding and the resolution of any remaining base-conflicts occurring within homopolymeric tracts. Artemis (66) was employed to inspect the results of the ORF prediction and its associated BLASTP results, which was used for a manual editing effort.

Multiple alignment of nucleotide sequences of the newly sequenced phages isolated in this study and those of previously sequenced members representing the four known *S.*

thermophilus phage groups [DT1 (cos) (67), TP-J34 (pac) (49), 5093 (5093) (9) and 9871 (987) (10)] were performed using ClustalW software. The alignment was employed to generate an unrooted phylogenetic tree using the "itol" software (http://itol.embl.de/) applying the neighbor-joining method.

Electron Microscopy

Phages STP1 and MM25 were selected as representatives of the phage collection for imaging and fresh, high titre lysates (at least 10^9 pfu.ml^{-1}) were produced and diluted 1:100 in SM buffer (68) before imaging. Adsorption of CsCl-purified phages to freshly prepared carbon film floated from a freshly coated mica sheet and negative staining with 2 % (w/v) uranyl acetate were performed as described previously (69). The film was picked up with a 400-mesh copper grid (Agar Scientific, Essex, UK), and specimens were examined with a Tecnai 10 transmission electron microscope (FEI, Eindhoven, The Netherlands) operated at an acceleration voltage of 80 kV.

Conflict of interest

The authors confirm that there is no conflict of interest associated with this work.

Acknowledgments

J.M. is supported by a Starting Investigator Research Grant (SIRG) (Ref. No. 15/SIRG/3430) funded by Science Foundation Ireland (SFI). D.v.S. is supported by a Principal Investigator award (Ref. No. 13/IA/1953) through SFI.

The authors wish to acknowledge the Irish cheese factory for kindly providing the samples for the phage screening performed in this study.

724

725 **References**

726

727 1. Auclair J, Accolas JP. 1983. Use of thermophilic lactic starters in the dairy industry. Antonie van
728 Leeuwenhoek 49:313-326.

729 2. Giraffa G, Paris A, Valcavi L, Gatti M, Neviani E. 2001. Genotypic and phenotypic heterogeneity
730 of *Streptococcus thermophilus* strains isolated from dairy products. J Appl Microbiol 91:937-943.

731 3. Duplessis M, Levesque, C. M. and S. Moineau. 2006. Characterization of *Streptococcus*
732 *thermophilus* host range phage mutants. Appl Environ Microbiol 72:3036-3041.

733 4. Quiberoni A, Stiefel, J. I., Reinheimer, J. A. 2000. Characterization of phage receptors in
734 *Streptococcus thermophilus* using purified cell walls obtained by a simple protocol. J Appl
735 Microbiol 89:1059-1065.

736 5. Fujisawa H, Morita M. 1997. Phage DNA packaging. Genes Cells 2:537-45.

737 6. Binetti AG, Del Rio, B., Martin, M. C., Alvarez, M. A. 2005. Detection and characterization of
738 *Streptococcus thermophilus* bacteriophages by use of the antireceptor gene sequence. Appl
739 Environ Microbiol 71:6096-6103.

740 7. Le Marrec C, van Sinderen D, Walsh L, Stanley E, Vlegels E, Moineau S, Heinze P, Fitzgerald G,
741 Fayard B. 1997. Two groups of bacteriophages infecting *Streptococcus thermophilus* can be
742 distinguished on the basis of mode of packaging and genetic determinants for major structural
743 proteins. Appl Environ Microbiol 63:3246-3253.

744 8. del Rio B, Binetti AG, Martin MC, Fernandez M, Magadan AH, Alvarez MA. 2007. Multiplex
745 PCR for the detection and identification of dairy bacteriophages in milk. Food Microbiol 24:75-
746 81.

747 9. Mills S, Griffin C, O'Sullivan O, Coffey A, McAuliffe OE, Meijer WC, Serrano LM, Ross RP.
748 2011. A new phage on the 'Mozzarella' block: Bacteriophage 5093 shares a low level of homology
749 with other *Streptococcus thermophilus* phages. Int Dairy J 21:963-969.

10. McDonnell B, Mahony J, Neve H, Hanemaaijer L, Noben JP, Kouwen T, van Sinderen D. 2016. Identification and analysis of a novel group of bacteriophages infecting the lactic acid bacterium *Streptococcus thermophilus*. Appl Environ Microbiol 82:5153-165.

11. Mahony J, Oliveira J, Collins B, Hanemaaijer L, Lugli GA, Neve H, Ventura M, Kouwen TR, Cambillau C, van Sinderen D. 2017. Genetic and functional characterisation of the lactococcal P335 phage-host interactions. BMC Genomics 18:146.

12. Mahony J, Deveau H, Mc Grath S, Ventura M, Canchaya C, Moineau S, Fitzgerald GF, van Sinderen D. 2006. Sequence and comparative genomic analysis of lactococcal bacteriophages jj50, 712 and P008: evolutionary insights into the 936 phage species. FEMS Microbiol Lett 261:253-261.

13. Szczepanska AK, Hejnowicz MS, Kolakowski P, Bardowski J. 2007. Biodiversity of *Lactococcus lactis* bacteriophages in Polish dairy environment. Acta Biochim Pol 54:151-158.

14. Raiski A, Belyasova N. 2009. Biodiversity of *Lactococcus lactis* bacteriophages in the Republic of Belarus. Int J Food Microbiol 130:1-5.

15. Murphy J, Royer B, Mahony J, Hoyles L, Heller K, Neve H, Bonestroo M, Nauta A, van Sinderen D. 2013. Biodiversity of lactococcal bacteriophages isolated from 3 Gouda-type cheese-producing plants. J Dairy Sci 96:4945-4957.

16. Suarez V, Moineau S, Reinheimer J, Quiberoni A. 2008. Argentinean *Lactococcus lactis* bacteriophages: genetic characterization and adsorption studies. J Appl Microbiol 104:371-379.

17. Verreault D, Gendron L, Rousseau GM, Veillette M, Masse D, Lindsley WG, Moineau S, Duchaine C. 2011. Detection of airborne lactococcal bacteriophages in cheese manufacturing plants. Appl Environ Microbiol 77:491-497.

18. Moineau S, Pandian S, Klaenhammer TR. 1993. Restriction/Modification systems and restriction endonucleases are more effective on lactococcal bacteriophages that have emerged recently in the dairy industry. Appl Environ Microbiol 59:197-202.

19. Quiberoni A, Tremblay D, Ackermann HW, Moineau S, Reinheimer JA. 2006. Diversity of *Streptococcus thermophilus* phages in a large-production cheese factory in Argentina. J Dairy Sci 89:3791-3799.

20. Szymczak P, Janzen T, Neves AR, Kot W, Hansen LH, Lametsch R, Neve H, Franz CMAP, Vogensen FK. 2017. Novel variants of *Streptococcus thermophilus* bacteriophages are indicative of genetic recombination among phages from different bacterial species. Appl Environ Microbiol 83(5).

21. Achigar R, Magadan AH, Tremblay DM, Pianzzola MJ, Moineau S. 2017. Phage-host interactions in *Streptococcus thermophilus*: Genome analysis of phages isolated in Uruguay and ectopic spacer acquisition in CRISPR array. Sci Rep 7.

22. Rousseau GM, Moineau S. 2009. Evolution of *Lactococcus lactis* phages within a cheese factory. Appl Environ Microbiol 75:5336-5344.

23. Madera C, Monjardin C, Suarez JE. 2004. Milk contamination and resistance to processing conditions determine the fate of *Lactococcus lactis* bacteriophages in dairies. Appl Environ Microbiol 70:7365-7671.

24. Quiberoni A, Guglielmotti DM, Reinheimer JA. 2003. Inactivation of *Lactobacillus delbrueckii* bacteriophages by heat and biocides. Int J Food Microbiol 84:51-62.

25. Deveau H, Barrangou R, Garneau JE, Labonte J, Fremaux C, Boyaval P, Romero DA, Horvath P, Moineau S. 2008. Phage response to CRISPR-encoded resistance in *Streptococcus thermophilus*. J Bacteriol 190:1390-1400.

26. Paez-Espino D, Sharon I, Morovic W, Stahl B, Thomas BC, Barrangou R, Banfield JF. 2015. CRISPR immunity drives rapid phage genome evolution in *Streptococcus thermophilus*. MBio 6 (2).

27. Horvath P, Romero DA, Coute-Monvoisin AC, Richards M, Deveau H, Moineau S, Boyaval P, Fremaux C, Barrangou R. 2008. Diversity, activity, and evolution of CRISPR loci in *Streptococcus thermophilus*. J Bacteriol 190:1401-1412.

28. Binetti AG, Reinheimer JA. 2000. Thermal and chemical inactivation of indigenous *Streptococcus thermophilus* bacteriophages isolated from Argentinian dairy plants. J Food Prot 63:509-515.

29. Hayes S, Murphy J, Mahony J, Lugli GA, Ventura M, Noben JP, Franz CM, Neve H, Nauta A, Van Sinderen D. 2017. Biocidal inactivation of *Lactococcus lactis* bacteriophages: Efficacy and targets of commonly used sanitizers. Front Microbiol 8:107.

30. Wagner N, Brinks E, Samtlebe M, Hinrichs J, Atamer Z, Kot W, Franz C, Neve H, Heller KJ. 2017. Whey powders are a rich source and excellent storage matrix for dairy bacteriophages. Int J Food Microbiol 241:308-317.

31. Lucchini S, Desiere F, Brussow H. 1999. Comparative genomics of *Streptococcus thermophilus* phage species supports a modular evolution theory. J Virol 73:8647-8656.

32. Desiere F, Lucchini S, Brussow H. 1998. Evolution of *Streptococcus thermophilus* bacteriophage genomes by modular exchanges followed by point mutations and small deletions and insertions. Virology 241:345-356.

33. McDonnell B, Mahony J, Hanemaaijer L, Neve H, Noben JP, Lugli GA, Ventura M, Kouwen TR, van Sinderen D. 2017. Global survey and genome exploration of bacteriophages infecting the lactic acid bacterium *Streptococcus thermophilus*. Front Microbiol 8:1754.

34. Kenny JG, McGrath S, Fitzgerald GF, van Sinderen D. 2004. Bacteriophage Tuc2009 encodes a tail-associated cell wall-degrading activity. J Bacteriol 186:3480-3491.

35. Stockdale SR, Mahony J, Courtin P, Chapot-Chartier MP, van Pijkeren JP, Britton RA, Neve H, Heller KJ, Aideh B, Vogensen FK, van Sinderen D. 2013. The lactococcal phages Tuc2009 and TP901-1 incorporate two alternate forms of their tail fiber into their virions for infection specialization. J Biol Chem 288:5581-5590.

36. Guglielmotti DM, Deveau H, Binetti AG, Reinheimer JA, Moineau S, Quiberoni A. 2009. Genome analysis of two virulent *Streptococcus thermophilus* phages isolated in Argentina. Int J Food Microbiol 136:101-109.

826 37. Stanley E, Fitzgerald GF, Le Marrec C, Fayard B, van Sinderen D. 1997. Sequence analysis and
827 characterization of phi O1205, a temperate bacteriophage infecting *Streptococcus thermophilus*
828 CNRZ1205. Microbiology 143 (Pt 11):3417-3429.
829 38. Lucchini S, Desiere F, Brussow H. 1999. The genetic relationship between virulent and temperate
830 *Streptococcus thermophilus* bacteriophages: whole genome comparison of cos-site phages Sfi19
831 and Sfi21. Virology 260:232-243.
832 39. Veesler D, Cambillau C. 2011. A common evolutionary origin for tailed-bacteriophage functional
833 modules and bacterial machineries. Microbiol Mol Biol Rev 75:423-433.
834 40. Duplessis M, Moineau S. 2001. Identification of a genetic determinant responsible for host
835 specificity in *Streptococcus thermophilus* bacteriophages. Mol Microbiol 41:325-336.
836 41. Bebeacua C, Tremblay D, Farenc C, Chapot-Chartier MP, Sadovskaya I, van Heel M, Veesler D,
837 Moineau S, Cambillau C. 2013. Structure, adsorption to host, and infection mechanism of virulent
838 lactococcal phage p2. J Virol 87:12302-12312.
839 42. Spinelli S, Campanacci V, Blangy S, Moineau S, Tegoni M, Cambillau C. 2006. Modular
840 structure of the receptor binding proteins of *Lactococcus lactis* phages. The RBP structure of the
841 temperate phage TP901-1. J Biol Chem 281:14256-14262.
842 43. Dieterle ME, Spinelli S, Sadovskaya I, Piuri M, Cambillau C. 2017. Evolved distal tail
843 carbohydrate binding modules of *Lactobacillus* phage J-1: a novel type of anti-receptor
844 widespread among lactic acid bacteria phages. Mol Microbiol 104:608-620.
845 44. Legrand P, Collins B, Blangy S, Murphy J, Spinelli S, Gutierrez C, Richet N, Kellenberger C,
846 Desmyter A, Mahony J, van Sinderen D, Cambillau C. 2016. The atomic structure of the phage
847 Tuc2009 baseplate tripod suggests that host recognition involves two different carbohydrate
848 binding modules. MBio 7:e01781-15.
849 45. Li X, Koc C, Kuhner P, Stierhof YD, Krismer B, Enright MC, Penades JR, Wolz C, Stehle T,
850 Cambillau C, Peschel A, Xia G. 2016. An essential role for the baseplate protein Gp45 in phage
851 adsorption to *Staphylococcus aureus*. Sci Rep 6:26455.

852 46. Koc C, Xia G, Kuhner P, Spinelli S, Roussel A, Cambillau C, Stehle T. 2016. Structure of the
853 host-recognition device of *Staphylococcus aureus* phage varphi11. Sci Rep 6:27581.
854 47. Mills S, Griffin C, O'Sullivan O, Coffey A, McAuliffe OE, Meijer WC, Serrano LM, Ross RP.
855 2011. A new phage on the 'Mozzarella' block: Bacteriophage 5093 shares a low level of homology
856 with other *Streptococcus thermophilus* phages. International Dairy Journal 21:963-969.
857 48. Zinno P, Janzen T, Bennedsen M, Ercolini D, Mauriello G. 2010. Characterization of
858 *Streptococcus thermophilus* lytic bacteriophages from mozzarella cheese plants. Int J Food
859 Microbiol 138:137-144.
860 49. Neve H, Zenz KI, Desiere F, Koch A, Heller KJ, Brussow H. 1998. Comparison of the lysogeny
861 modules from the temperate *Streptococcus thermophilus* bacteriophages TP-J34 and Sfi21:
862 implications for the modular theory of phage evolution. Virology 241:61-72.
863 50. Mavrich TN, Hatfull GF. 2017. Bacteriophage evolution differs by host, lifestyle and genome. Nat
864 Microbiol 2:17112.
865 51. Koberg S, Mohamed MD, Faulhaber K, Neve H, Heller KJ. 2015. Identification and
866 characterization of cis- and trans-acting elements involved in prophage induction in *Streptococcus*
867 *thermophilus* J34. Mol Microbiol 98:535-552.
868 52. Mahony J, Stockdale SR, Collins B, Spinelli S, Douillard FP, Cambillau C, van Sinderen D. 2016.
869 *Lactococcus lactis* phage TP901-1 as a model for *Siphoviridae* virion assembly. Bacteriophage
870 6:e1123795.
871 53. Mahony J, Randazzo W, Neve H, Settanni L, van Sinderen D. 2015. Lactococcal 949 group
872 phages recognize a carbohydrate receptor on the host cell surface. Appl Environ Microbiol
873 81:3299-3305.
874 54. Chopin A, Bolotin A, Sorokin A, Ehrlich SD, Chopin M. 2001. Analysis of six prophages in
875 *Lactococcus lactis* IL1403: different genetic structure of temperate and virulent phage
876 populations. Nucleic Acids Res 29:644-651.
877 55. Lillehaug D. 1997. An improved plaque assay for poor plaque-producing temperate lactococcal
878 bacteriophages. J Appl Microbiol 83:85-90.

879 56. Dupont K, Vogensen FK, Josephsen J. 2005. Detection of lactococcal 936-species bacteriophages
880 in whey by magnetic capture hybridization PCR targeting a variable region of receptor-binding
881 protein genes. J Appl Microbiol 98:1001-1009.
882 57. Staden R, Judge DP, Bonfield JK. 2001. Sequence assembly and finishing methods. Methods
883 Biochem Anal 43:303-322.
884 58. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic
885 gene recognition and translation initiation site identification. BMC Bioinformatics 11:119.
886 59. Gish W, States DJ. 1993. Identification of protein coding regions by database similarity search.
887 Nat Genet 3:266-272.
888 60. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped
889 BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids
890 Res 25:3389-3402.
891 61. Lugli GA, Milani C, Mancabelli L, van Sinderen D, Ventura M. 2016. MEGAnnotator: a user-
892 friendly pipeline for microbial genomes assembly and annotation. FEMS Microbiol Lett 363.
893 62. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M,
894 Moxon S, Sonnhammer EL, Studholme DJ, Yeats C, Eddy SR. 2004. The Pfam protein families
895 database. Nucleic Acids Res 32:D138-141.
896 63. Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH,
897 Geer LY, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Jackson JD, Ke Z, Lanczycki CJ, Lu F,
898 Marchler GH, Mullokandov M, Omelchenko MV, Robertson CL, Song JS, Thanki N, Yamashita
899 RA, Zhang D, Zhang N, Zheng C, Bryant SH. 2011. CDD: a Conserved Domain Database for the
900 functional annotation of proteins. Nucleic Acids Res 39:D225-229.
901 64. Soding J, Biegert A, Lupas AN. 2005. The HHpred interactive server for protein homology
902 detection and structure prediction. Nucleic Acids Res 33:W244-248.
903 65. Lowe TM, Eddy SR. 1997. tRNAscan-SE: A program for improved detection of transfer RNA
904 genes in genomic sequence. Nucleic Acids Research 25:955-964.

66. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B. 2000. Artemis: sequence visualization and annotation. Bioinformatics 16:944-945.

67. Tremblay DM, Moineau S. 1999. Complete genomic sequence of the lytic bacteriophage DT1 of *Streptococcus thermophilus*. Virology 255:63-76.

68. Murphy J, Bottacini F, Mahony J, Kelleher P, Neve H, Zomer A, Nauta A, van Sinderen D. 2016. Comparative genomics and functional analysis of the 936 group of lactococcal *Siphoviridae* phages. Sci Rep 6:21345.

69. Deasy T, Mahony J, Neve H, Heller KJ, van Sinderen D. 2011. Isolation of a virulent *Lactobacillus brevis* phage and its application in the control of beer spoilage. J Food Prot 74:2157-2161.

932 **Table 1.** Details of the samples tested in this study and the phages isolated between 2006 and 2016
933

Year[♣]	# samples	*Total # tests performed (# samples × 52 bacterial strains)	# phage positive samples	# phages isolated	Isolate genotype (# isolates)
Factory A whey samples					
2006	105	n/a	68	14	STP1 (14)
2007	98	n/a	58	12	STP1 (12)
2008	97	n/a	62	16	STP1 (15), STP2 (1)
2009	102	n/a	59	14	STP1 (14)
2010	101	n/a	67	12	STP1 (10), STP2 (2)
2011	109	n/a	65	20	STP1 (16), STP2 (4)
2012	115	n/a	65	16	STP1 (16)
2013	106	n/a	68	14	STP1 (12), STP2 (2)
2014	98	n/a	66	12	STP1 (12)
2015	101	n/a	69	16	STP1 (6), STP2 (2), A0 (1), B0 (1), C0 (1), 9B4 (1), 16B8 (2), 31B4 (2)
2016	104	n/a	64	14	STP1 (12), 7T (2)
TOTAL	1,136	n/a	711 (63 %)	160	
Externally-acquired whey samples					
2006	94	4,888	62	10	STP1 (10)
2007	98	5,096	60	12	STP1 (12)
2008	96	4,992	47	12	STP1 (12)
2009	92	4,784	54	12	STP1 (10), STP2 (2)
2010	100	5,200	59	14	STP1 (14)
2011	96	4,992	69	12	STP1 (9), STP2 (3)
2012	290	15,080	207	14	STP1 (4), STP2 (5), B5 (5)
2013	306	15,912	223	14	STP1 (6), B5 (8)
2014	295	15,340	201	14	B5 (10), MM25 (2), M19 (2)
2015	305	15,860	244	16	STP1 (2), 9A (2), L5A1 (2), 7A5 (4), 7T (6)
2016	271	14,092	164	16	A0 (2), B5 (2), STP2 (4), V2 (5), R1 (3)
TOTAL	2,043	106,236	1390 (68 %)	146	

934 [♣]The sampling period for each year was between March and early July.

935 *This refers only to externally-acquired samples as Factory A-derived samples were all pre-screened on
936 *S. thermophilus* P1 and only the isolated propagated phages were tested against the panel of 52 *S.*
937 *thermophilus* strains.

938

Table 2. Host range highlighting the average E.O.P. of phage isolates relative to the primary host* (bold) and details of the origin of the isolates. Shaded areas indicate strains that were insensitive to infection by the isolated phages.

Phage isolate	Year of isolation	Source	P1	P2	P3	P4	P5	P6	CNRZ447	STR1
STP1	2006	Factory	**1**	1.1×10^{-5}	2.6×10^{-3}					
STP2	2008	Factory	**1**	1						
31B4	2015	Factory	**1**	2.6×10^{-2}		1		1.6×10^{-5}		
16B8	2015	Factory	**1**	6.8×10^{-4}		0.7				
9B4	2015	Factory	**1**	1	6.8×10^{-4}			8.2×10^{-5}		
A0	2015	Factory	**1**	2.4×10^{-2}				1.5×10^{-4}		
B0	2015	Factory	**1**	1.3×10^{-4}						
C0	2015	Factory	**1**		0.5					1
B5	2012	External	**1**	1.2×10^{-1}						
MM25	2014	External					**1**			
M19	2014	External		**1**					3.3×10^{-2}	
9A	2015	External	**1**	1.5×10^{-2}				1.0×10^{-4}		
L5A1	2015	External	**1**	1.7×10^{-1}						
7A5	2015	External	**1**	8.5×10^{-2}		6.7×10^{-3}				
7T	2015	External	**1**	1						
V2	2016	External	**1**			2.9×10^{-2}				
R1	2016	External	0.3		0.5					**1**

* The host range of each of the phages was assessed against a panel of 52 *S. thermophilus* strains. Only those strains that were sensitive to infection by one or more of the phages is presented in the table above.

Table 3. Order of appearance and source of first isolates of lineage 1 and 2 phages in factory A

Phage isolated	Date of isolation	Location of first appearance	Re-occurrence in factory	Subsequent factory-isolated related phages
STP1 (Lineage 1)	March 2006	Factory A	2007-2016	STP2 (May 2008), A0, B0, C0 (April 2015)
7T (Lineage 2)	March 2015	External factory	2016	9B4, 16B8, 31B4 (June 2015), 7T (2016)

971 **Figure legends**

972

973 **Fig. 1.** Unrooted phylogenetic tree of the 17 sequenced isolates and a representative *cos* (DT1),
974 *pac* (TP-J34), 5093 (5093) and 987 (9871) groups included as references. All phage isolates group
975 most closely to DT1 while distinct subgroups or lineages (L) of the phage isolates are also
976 highlighted in the tree (L1- L4). Factory-derived phage isolates are indicated in black text;
977 externally-derived phage isolates are indicated in green text and representatives of the four *S.*
978 *thermophilus* groups (*cos, pac*, 5093 and 987) are indicated in blue text.

979

980 **Fig. 2.** Schematic representation of the genomes of lineage 1 (STP1, STP2, C0, A0, 7A5, B5, B0,
981 L5A1) and lineage 2 (7T, 9B4, 31B4, 16B8, R1, V2) phages. Arrows (indicating protein-encoding
982 regions) joined by shaded boxes indicate genetic regions of similarity with the black shading
983 indicating 90-100 % aa identity; dark grey indicating 80-89 %; light grey indicating 50-79 % and
984 off-white indicating 30-49 % aa identity. Arrows of the same colour represent genes with a similar
985 function. Those arrows coloured in grey are indicative of genes encoding proteins of unknown
986 function. The predicted functions of the encoded proteins are presented above the relevant arrows,
987 where known and the functional modules are presented below the schematic. The major region of
988 divergence between 7T (the first lineage 2 isolate) and STP1 (first lineage 1 isolate) is highlighted
989 in the 7T genome representation by a red box. Similarly, regions of genetic novelty associated
990 with isolates B5, 7A5 and L5A1 are highlighted in red boxes. All lineage 2 phage genomes
991 possess a genomic region with greater than 90 % identity. The lineage (L1/2) and source (factory-
992 F or external-E) are also presented on the left side of the figure.

993

994

995 **Fig. 3.** Schematic representation of the genomes of lineage 3 (MM25 and 9A) and lineage 4 (M19)
996 phages and their comparison to the first isolate of the study (STP1, lineage 1). Arrows (indicating
997 protein-encoding regions) joined by shaded boxes indicate genetic regions of similarity with the

black shading indicating 90-100 % aa identity; dark grey indicating 80-89 %; light grey indicating 50-79 % and off-white indicating 30-49 % aa identity. Arrows of the same colour represent genes with a similar function. Those arrows coloured in grey are indicative of genes encoding proteins of unknown function. The predicted functions of the encoded proteins are presented above the relevant arrows, where known and the functional modules are presented below the schematic. The lineage (L3/4) and source (factory-F or external-E) are also presented on the left side of the figure.

Fig. 4. Top: Unrooted phylogenetic tree of the Tal-RBPs of the 17 sequenced phages highlighting the disparity of the Tal-RBPs of the lineage 1/2 (left) and 3/4 phages (right). Bottom: Schematic depicting the organisation of the Tal-RBPs of the sequenced phages using representatives of the group. All Tal-RBPs possess a conserved N-terminal ~400 aa Tal domain (blue). 9A (lineage 3 phage) is the sole phage encoding a Tal-RBP with two BppA-like (5E7T_B) CBDs (purple) and these constitute the regions described as VR1 and VR2. VR1, where present, is always flanked by collagen repeat motifs (red). The VR2 region may/may not incorporate a 5E7T_B BppA CBD (purple) and/or may represent a distinct CBD, e.g. Igu1_A domain (yellow), or a CBD of, as yet, unidentified structure. STP1 is representative of the Tal-RBPs of lineage 1 and 2 phages, which all share a similar size and architecture. MM25 is representative of the lineage 4 phages.

Fig. 5. Representative electron micrographs of STP1 (panel A) and MM25 (panel B). Both phages display long tails with protruding feather-like appendages from the tail tip.